DOCUMENT RESUME

ED 472 886 IR 058 591

AUTHOR	Chan, Lois Mai; Zeng, Marcia Lei
TITLE	Ensuring Interoperability among Subject Vocabularies and Knowledge Organization Schemes: A Methodological Analysis.
PUB DATE	2002-08-00
NOTE	9p.; In: Libraries for Life: Democracy, Diversity, Delivery. IFLA Council and General Conference: Conference Programme and Proceedings (68th, Glasgow, Scotland, August 18-24, 2002); see IR 058 549.
AVAILABLE FROM	For full text: http://www.ifla.org.
PUB TYPE	Information Analyses (070) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Information Retrieval; Information Systems; Internet; Knowledge Representation; Online Systems; *Vocabulary; World Wide Web
IDENTIFIERS	*Interoperability; Query Languages; Query Processing

ABSTRACT

The heterogeneous environment of information retrieval on the World Wide Web has brought the recognition for the need of interoperability among diverse systems to the fore. In subject retrieval, users encounter not only different vocabularies and schemes, but also different languages. As a result, there has been a flourish of projects in the last few years aimed at improving the interoperability among subject vocabularies and knowledge organization schemes, with some targeting different vocabularies and others focusing on different languages. This paper attempts to analyze the methods used in these projects. It begins with a brief overview and then examines in particular the approaches and methods used in recent efforts. (Contains 18 references.) (Author)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 472 886

68th IFLA Council and General Conference August 18-24, 2002

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

S. Koopman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Code Number: 008-122-E
Division Number: IV
Professional Group: *Classification and Indexing*
Joint Meeting with: -
Meeting Number: 122
Simultaneous Interpretation: -

Ensuring Interoperability among Subject Vocabularies and Knowledge Organization Schemes: a Methodological Analysis

Lois Mai Chan,
School of Library and Information Science, University of Kentucky
USA

Marcia Lei Zeng
School of Library and Information Science, Kent State University
Kent, USA

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Abstract:

The heterogeneous environment of information retrieval on the World Wide Web has brought the recognition for the need of interoperability among diverse systems to the fore. In subject retrieval, users encounter not only different vocabularies and schemes but also different languages. As a result, there has been a flourish of projects in the last few years aimed at improving the interoperability among subject vocabularies and knowledge organization schemes, with some targeting different vocabularies and others focusing on different languages. This paper attempts to analyze the methods used in these projects. It begins with a brief overview and then examines in particular the approaches and methods used in recent efforts.

IR058591

BEST COPY AVAILABLE

1 INTRODUCTION

In the open environment of the Internet and the Web, information resources are heterogeneous and have been indexed with different vocabularies and organized according to different schemes. How to achieve the best retrieval results in cross-domain searching has presented a particular challenge to the information profession. In information retrieval, users typically are not, nor should they need to be, aware of the behind-the-scenes mechanisms for matching their query terms to the vocabularies employed by various systems. The ideal approach would be to provide a "one-stop" seamless searching instead of requiring the user to search individual databases or collections separately. To enable such an approach, it is important to render the different knowledge organization systems, such as controlled vocabularies and classification schemes, interoperable within a single search apparatus.

2. REVIEW OF PROJECTS AND EXAMPLES OF INTEROPERABLE VOCABULARIES

Before we begin examining their methods, let us review briefly a number of recent efforts in achieving interoperability between and among different subject vocabularies (including both controlled and uncontrolled vocabularies) and knowledge organization systems. These include efforts at establishing interoperability among vocabularies in the same language or in different languages, among different classification schemes, and between controlled vocabularies and classification schemes. These efforts have led to the mapping and integration of existing knowledge organization systems, or the creation of new ones, for information sharing in a networked environment. The projects have varied in both the targets for mapping and the methods used in achieving their aims. The projects we discuss, organized by similarity of task, are the following:

2.1 Among Controlled Vocabularies in the Same Language

1. Between *Library of Congress subject headings* (LCSH) and *Medical subject headings* (MeSH) - Northwestern University (Olson, 2001).
2. Among different controlled vocabularies - H.W. Wilson Company (Kuhr, 2001)
3. Among different German thesauri that are used to index mathematics and physics as well as social science literature– CARMEN (Content Analysis, Retrieval, Metadata: Effective Networking) (CARMEN WP12, 2000)

2.2 Among Multiple Subject Vocabularies in Different Languages and Classification systems

1. Among thesauri, classification systems, coding systems, and lists of controlled terms in biomedical fields – UMLS (Unified Medical Language System) Metathesaurus (National Library of Medicine, 2001)
2. Among distributed services employing different indexing vocabularies used by various communities such as archives, the Further and Higher Education sectors, libraries, museums, the National Grid for Learning, and the Resource Discovery Network, etc. - HILT (High-Level Thesaurus Project). (HILT, 2000; Nicholson, Wake and Currier, 2001a)
3. Among the "Entry vocabularies" used by systems (e.g., indexes to *BIOSIS Concept Codes*, *INSPEC Thesaurus*, *U.S. Patent and Trademark Office Patent Classification, etc.)* in order to map them to "Query vocabularies" entered in a search. – University of California Berkeley DARPA Unfamiliar Metadata Project (Buckland et al., 1999).
4. Among local class schemes to a common scheme (DDC (*Dewey Decimal Classification*) - Renardus project (Koch, Neuroth, and Day, 2001)
5. Among four controlled vocabularies and schemes: *Polish Thematic Classification* (PTC), descriptors based on the *Thesaurus of Common Topics* (TCT), *Universal Decimal Classification* (UDC), and

Subject-Heading Language (SHL) of the National Library in Warsaw – Polish Project (Scibor and Tomasik-Beck, 1994).

6. Among controlled vocabularies used by four national libraries' catalogs in three languages: English, French, and German - MACS (Multilingual Access to Subjects) (Freyre and Naudi, 2001).
7. Among vocabularies for a multilingual database about the French heritage – Merimee (See statistics reported in Doerr, 2001.)

2.3 Between a Controlled Vocabulary and a Universal Classification System

1. Between LCSH and LCC (*Library of Congress Classification*) –*Classification Plus* (a CD-ROM product) and *Classification Web* (a web-based interface under development), Library of Congress
2. Between LCSH and DDC (Vizine-Goetz, 1996)
3. Between UDC and GFSH (*General Finish Subject headings*) (Himanka and Vesa, 1992)

2.4 Between Classification Systems

1. Between MSC (the American Mathematical Society (AMS) *Mathematics Subject Classification*) and Schedule 510 in DDC - State University of New York in Albany, New York. (Iyer and Giguere, 1995).
2. Between SAB (*Klassifikationssystem för svenska bibliotek)* and DDC – Swedish Royal Library (IFLA, 2001:34)

2.5 New System for Different Languages

The HEREIN Project (The European information network on cultural heritage policies) produced an interlingua, a thesaurus consisting of terms derived from reports on cultural heritage policies in Europe. It was created with no direct reference to the terms or to the structure of any pre-existing thesaurus. - The HEREIN Project (http://www.european-heritage.net/en/index.html, click Thesaurus)

3 METHODS USED FOR ACHIEVING AND IMPROVING INTEROPERABILITY

The concern for vocabulary compatibility is not new. Long before the advent of the electronic age, library and information professionals had explored and employed various methods to reduce conflict between different vocabularies that were used in the same system. Earlier methods relied almost completely on intellectual efforts. As advanced computerized process methods for achieving or improving interoperability emerged, computer technology began to be used to fully benefit from the networked environment. The following section lists both conventional and new methods that have become widely accepted.

1. Derivation/Modeling - A specialized or simpler vocabulary is developed with an existing, more comprehensive vocabulary as a starting point or model.
2. Translation/Adaptation – A controlled vocabulary is developed which consists of terms translated from one in a different language with or without modification.
3. Mapping (intellectual) – A mapping system is developed which consists basically of establishing equivalents between terms in different controlled vocabularies or between verbal terms and classification numbers. Such mapping generally requires a great deal of intellectual effort.

4. Mapping (computer-aided) – A mapping system is developed which relies partly or heavily on computer technology.
5. Linking – A list is developed of terms that linked with other terms that are not conceptual equivalents but are closely related linguistically. Such links have been found to enhance retrieval results.
6. Switching – A switching language or scheme is developed which serves as an intermediary for moving among equivalent terms in different vocabularies.

4 METHODS USED IN THE LINK STORAGE AND MANAGEMENT

Once the mapping is established, a device is needed for storing and maintaining the links to manage the large number of indexing terms and their complex relationships that result. For this purpose, several options have been explored and used:

1. Authority records - Special fields in authority formats may be used to store the links.
2. Concordances - The elaboration of concordances requires the discerning of one master vocabulary/scheme and of one or more target vocabularies/schemes.
3. Semantic network - A semantic network, also called a semantic web, consists of an organized structure serving as the "spine" or backbone. Each unit in the network represents a concept around which a cluster of equivalent terms from different vocabularies is identified and stored.

5. DISCUSSION

A number of common issues have emerged in our analysis of the methods used in the many projects discussed above.

5.1 General Issues in Mapping

5.1.1 Mapping multilingual vocabularies.

At the heart of multilingual subject vocabulary is mapping or establishing equivalence. One-to-one relationships between terms in different vocabularies and different languages are ideal matches, but are often elusive. Different linguistic expressions for the same concept, different degrees of specificity, and polysemous terms are some of the difficulties facing those attempting to map vocabularies and those creating multilingual or multi-disciplinary vocabularies. The complex requirements and processes of matching terms that are often imprecise have an impact on the following aspects of vocabulary mapping (Koch, Neuroth, and Day, 2001): browsing structure, display, depth, non-topical classes, and the trade-off between consistency, accuracy and usability. Various levels of mapping/linking can co-exist in the same project, such as those identified by the MACS project: terminological level (subject heading), semantic level (authority record), and syntactic level (application) (Freyre and Naudi, 2001).

5.1.2 Integrating the views of different cultures.

Under the assumption that all languages are equal in a concordance, there exists a question of whether the views of a particular culture that are expressed through a controlled vocabulary or a classification can be appropriately transferred to those in a different culture in the process of mapping. Hudon (1997) noticed the following problems associated with multilingual systems:

1) that of stretching a language to make it fit a foreign conceptual structure to the point where it becomes barely recognizable to its own speakers;
2) that of transferring a whole conceptual structure from one culture to another whether it is appropriate or not; and
3) that of translating literally terms from the source language into meaningless expressions in the target language, etc.

BEST COPY AVAILABLE

She summarizes the issues involved as: management issues, linguistic/semantic issues, and technology-related issues.

5.1.3 Mapping systems with different structures

There are basic differences in the terms of the macro-structures of controlled vocabularies and classification systems. Thesauri constructed by following ISO 2788 and other national standards ensure that the structure and "grammar" of such a vocabulary stay consistent or compatible. The construction of subject headings and classification schemes, on the other hand, has been guided by existing patterns or examples. Chances are, if there are ten different universal systems, there will be ten different guidelines. As a result, knowledge organization systems differ from one another in their structure, semantic, lexical, and notation or entry features. (Iyer and Giguere, 1995). For example, they may cover different subject domains, or with different scope and coverage; they may have semantic differences that are caused by variations in conceptual structuring; their levels of specificity and the use of terminology may vary; and the syntactic features, such as the word order of terms and the choice of reserved heading use, may be different.

These incompatibilities have presented problems for any mapping effort from the beginning. For example, establishing concordance or translation between a thesaurus and a classification or among various systems sometimes becomes impossible or extremely challenging. This is especially true when the target system has a higher level of specificity than the source system or other systems.

5.2 Methodological Options

For projects aiming at establishing interoperability between or among selected knowledge organization systems in order to meet user's new requirements in the networked environment, one major decision that needs to be made is the choice of an appropriate method. The first complex question to be answered is: to integrate, map, or create a new system? The options are similar to those suggested by Riesthuis (2001) with regard to different approaches to creating multilingual thesauri:

1. translation
2. merging
3. creating from scratch

Within each of these approaches are multiple possibilities, as suggested by the HILT researchers in a two-dimensional grid. (Nicholson, Wake and Currier, 2001b).
These researchers propose three basic options:

- Using or creating a single scheme (LCSH, UNESCO, DDC-based, UDC-based, entirely new);
- Mapping existing schemes (LCSH, UNESCO, DDC-based, UDC-based);
- Mapping existing schemes in the short term, leading to a single scheme in the long term.

Based on the options listed above, additional considerations can be applied:

- additional thesaurus structure;
- new subject specific micro-thesauri;
- mapping among existing domain specific micro-thesauri;
- multilingual capability;
- community control;
- machine-assisted methods;
- AI-assisted methods;

- user training;
- flexible facilities to aid users;
- user mind maps;
- consistency in term application ensured via training and monitoring;
- trained librarians to help the user optimize retrieval

The choice of the basic approach plus any combinations of the considerations may bring various end products and require different amount of time and resources. Any method and combinations with other processes may have pros and cons. It is necessary to conduct a comprehensive research and to identify potential problems when a particular method is employed.

6. CONCLUSION

1. What have we learned from the projects?
2. What issues are still outstanding?
3. What is needed, in terms of intellectual and technical approaches, to move the field forward?

From the examples introduced in this paper, we can summarize the following trends that form the mainstream:

1. The need for interoperability among knowledge organization systems is an unavoidable issue and process in today's networked environment.
2. Various methods have been used in achieving interoperability among knowledge organization systems. It may or may not be that a switching system will be needed. It may or may not be that building a concordance between or among the involved vocabularies may be the ideal situation. Or, equally possible, it may be that interoperability may be more effectively achieved through the subject authority records of various online systems.
3. While mapping vocabularies is still a largely intellectual effort, computer technology has been applied to assist in managing large files of subject data and in managing links. Higher levels of computerized mapping systems have also been subject to experimentation or tests. Both human mapping and computer-aided mapping will co-exist for a period of time to come.
4. Numerous projects for cross-language and cross-structure mapping have been initiated. These projects have identified and experimented with a variety of methods. It is safe to predict that there will be many more multilingual products and services, and many of them will involve multiple structured systems such as thesaurus, classification, subject headings, and index terms assigned to database records.

The need for reconciling different subject vocabularies in the networked environment is indisputable. Results from recent efforts in achieving interoperability among vocabularies of different sorts and in different languages are encouraging. The question remains: Have we fully exploited technological capabilities in our efforts to improve subject access to the myriads of resources now available in the networked environment?

ACKNOWLEDGEMENT

Grateful acknowledgement is made to the principal investigators of the interoperability projects discussed in this paper: Traugott Koch (Sweden and Denmark), Patricia Kuhr (USA), Martin Kunz (Germany), Max Naudi (France), Dennis Nicholson (UK), and Tony Olson (USA), who were generous in taking their time to answer our questions, provide details regarding their projects, or preview the paper.

REFERENCES

Buckland, M., et al. *(1999). Mapping entry vocabulary to unfamiliar metadata vocabularies,* D-Lib Magazine, 5(1). *http://www.dlib.org/dlib/january99/buckland/01buckland.html, (last accessed Feb. 5, 2002).*

CARMEN. WP12: Cross concordances of classifications and thesauri. *http://www.bibliothek.uni-regensburg.de/projects/carmen12/index.html.en* (last accessed Feb. 5, 2002)

Doerr, Martin. (2001) Semantic problems of thesaurus mapping. *Journal of Digital information*, 1 (8). http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr/#Nr.52

Freyre, Elisabeth and Max Naudi. (2001) MACS: Subject access across languages and networks. In *Subject Retrieval in a Networked Environment: Papers Presented at an IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, OCLC, Dublin, Ohio, USA, 14-16 August 2001.* Dublin, OH: OCLC.

HILT. (2000) *HILT: High-Level Thesaurus Project Proposal.* http://hilt.cdlr.strath.ac.uk/AboutHILT/proposal.html. (Last accessed Feb.5, 2002)

Himanka, Janne and Kautto Vesa. (1992) Translation of the Finish Abridged Edition of UDC into General Finish Subject Headings. *International Classification* 19(3):131-134.

Hudon, Michele. (1997) Multilingual thesaurus construction: integrating the views of different cultures in one gateway to knowledge concepts. *Knowledge Organization* 24(2): 84-91.

IFLA Section on Classification and Indexing. (2001) *Newsletter* Nr.24, December 2001.

Iyer, Hermalata and Mark Giguere. (1995). Towards designing an expert system to map mathematics classificatory structures. *Knowledge Organization* 22(3/4):141-147.

Koch, Traugott, Heike Neuroth, and Michael Day. (2001) Renardus: cross-browsing european subject gateways via a common classification system (DDC). In *Subject Retrieval in a Networked Environment: Papers Presented at an IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, OCLC, Dublin, Ohio, USA, 14-16 August 2001.* Dublin, OH: OCLC. http://www.lub.lu.se/~traugott/drafts/preifla-final.html (last accessed Feb.5, 2002)

Kuhr, Patricia S. (2001) Putting the world back together: mapping multiple vocabularies into a single thesaurus. In *Subject Retrieval in a Networked Environment: Papers Presented at an IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, OCLC, Dublin, Ohio, USA, 14-16 August 2001.* Dublin, OH: OCLC.

National Library of Medicine. (2001) *Fact Sheet: UMLS ® Metathesaurus ®* Last updated 2001. http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html. (last accessed Feb. 5, 2002)

Nicholson, Dennis and Susannah Wake. (2001a) HILT: Subject retrieval in a distributed environment. In *Subject Retrieval in a Networked Environment: Papers Presented at an IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, OCLC, Dublin, Ohio, USA, 14-16 August 2001.* Dublin, OH: OCLC.

BEST COPY AVAILABLE

Nicholson, Dennis, Susannah Wake, and Sarah Currier. (2001b) High-Level Thesaurus Project: investigating the problem of subject cross-searching and browsing between communities. In *Global Digital Library Development in the New Millemnnium: fertile ground for distributed cross-disciplinary collaboration*, edited by Ching-Chih Chen. Beijing: Tsinghua University Press, 2001.

Olson, Tony. (2001) Integrating LCSH and MeSH in information systems. In *Subject Retrieval in a Networked Environment: Papers Presented at an IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, OCLC, Dublin, Ohio, USA, 14-16 August 2001.* Dublin, OH: OCLC.

Riesthuis, Gerhard J.A. (2001) Information languages and multilingual subject access. In *Subject Retrieval in a Networked Environment: Papers Presented at an IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, OCLC, Dublin, Ohio, USA, 14-16 August 2001.* Dublin, OH: OCLC.

Scibor, Eugeniusz and Joanna Tomasik-Beck. (1994) On the establishment of concordances between indexing languages of universal or interdisciplinary scope (Polish Experiences). *Knowledge Organization* 21(4):203-212.

Vizine-Goetz, Diane. (1996) Classification Research at OCLC. *Annual Review of OCLC Research*, pp. 27-33.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").